JOURNAL OF APPLIED ORAL SCIENCE
www.fob.usp.br/revista or www.scielo.br/jaos
J Appl Oral Sci. *2007;15(6):480-5*

A NEW APPROACH FOR INTEREXAMINER RELIABILITY DATA ANALYSIS ON DENTAL CARIES CALIBRATION

Andréa Videira ASSAF[1], Elaine Pereira da Silva TAGLIAFERRO[2], Marcelo de Castro MENEGHIM[3], Cristiana TENGAN[2], Antonio Carlos PEREIRA[4], Gláucia Maria Bovi AMBROSANO[5], Fábio Luiz MIALHE[3]

1- DDS, MSc, PhD, Department of Community Services - Fluminense Federal University, Niterói, RJ, Brazil.
2- DDS, MSc, Graduate student, Department of Community Dentistry, Dental School of Piracicaba, State University of Campinas, Piracicaba, SP, Brazil.
3- DDS, MSc, PhD, Professor, Department of Community Dentistry, Dental School of Piracicaba, State University of Campinas, Piracicaba, SP, Brazil.
4- DDS, MPH, DrPH, Professor, Department of Community Dentistry, Dental School of Piracicaba, State University of Campinas, Piracicaba, SP, Brazil.
5- Agr.Eng., MSc, PhD, Professor, Department of Community Dentistry, Dental School of Piracicaba, State University of Campinas, Piracicaba, SP, Brazil.

Corresponding address: Prof. Dr. Antonio Carlos Pereira - Departamento de Odontologia Social - FOP - UNICAMP - Avenida Limeira 901 - Piracicaba, SP, Brasil - 13414-903 - Phone: +55-19-2106-5209. Fax: +55-19-2106-5218 - e-mail: apereira@fop.unicamp.br

Received: January 22, 2007- Modification: June 4, 2007 - Accepted: June 19, 2007

ABSTRACT

Objectives: a) to evaluate the interexaminer reliability in caries detection considering different diagnostic thresholds and b) to indicate, by using Kappa statistics, the best way of measuring interexaminer agreement during the calibration process in dental caries surveys. Methods: Eleven dentists participated in the initial training, which was divided into theoretical discussions and practical activities, and calibration exercises, performed at baseline, 3 and 6 months after the initial training. For the examinations of 6-7-year-old schoolchildren, the World Health Organization (WHO) recommendations were followed and different diagnostic thresholds were used: WHO (decayed/missing/filled teeth – DMFT index) and WHO + IL (initial lesion) diagnostic thresholds. The interexaminer reliability was calculated by Kappa statistics, according to WHO and WHO+IL thresholds considering: a) the entire dentition; b) upper/lower jaws; c) sextants; d) each tooth individually. Results: Interexaminer reliability was high for both diagnostic thresholds; nevertheless, it decreased in all calibration sections when considering teeth individually. Conclusion: The interexaminer reliability was possible during the period of 6 months, under both caries diagnosis thresholds. However, great disagreement was observed for posterior teeth, especially using the WHO+IL criteria. Analysis considering dental elements individually was the best way of detecting interexaminer disagreement during the calibration sections.

Uniterms: Epidemiology; Dental caries; Calibration; Data reliability.

INTRODUCTION

An important aspect of any research is the use of appropriate methodologies either to control or to reduce the effects of potential confounding factors. A matter of great concern that can influence the results of epidemiological studies of dental caries is the variation in disease diagnosis between two or more examiners (interexaminer error) and for the same examiner in two or more occasions (intraexaminer error).

Therefore, it is very important that data collection measures are standardized in order to minimize measurements variations[1]. The calibration process including the determination of reliability, with both previous and ongoing epidemiological survey, is a basic step to understand and standardize the examination criteria, and also to evaluate the interexaminer variability in order to ensure accurate results[15,17].

The assessment of reliability is the most employed measure in dental caries surveys during the examiners' calibration. Reliability is related to the extent to which examiners agree in their evaluations[16]. The most used measures to assess reliability in epidemiological studies of dental caries are the overall percentage of agreement and the Kappa statistics[11]. Kappa test is a measurement of reliability that takes into consideration the agreement among raters by chance, providing better evaluation of interexaminer disagreement during calibration processes[5].

The purposes of this study were: a) to evaluate interexaminer reliability in caries detection considering

different diagnostic thresholds and b) to indicate, by using Kappa statistics, which is the best way of measuring interexaminer agreement during the calibration process in dental caries surveys.

MATERIAL AND METHODS

Ethical aspects

The epidemiological examinations were initiated after approval of the study design by the Research Ethics Committee of the Dental School of Piracicaba, State University of Campinas, Brazil (protocol No. 068/2002). The volunteers' parents signed an informed consent form authorizing the enrollment of the children in the study.

Study Design

Dentists with previous experience in epidemiological surveys examined schoolchildren at baseline, 3 and 6 months after initial training, using two diagnostic thresholds on dental caries: WHO criteria, traditionally used in epidemiological surveys[17] and WHO+IL including the diagnostic of initial caries lesions (IL, white spot lesions), after being calibrated by a "gold standard" examiner, a dentist who routinely uses the WHO criteria for exams and had been previously trained and calibrated in IL diagnosis. Data were analyzed by Kappa statistics, considering distinct data approach.

Selection of Examiners and Schoolchildren

Eleven dentists with previous experience in epidemiological surveys of dental caries were invited to participate in this study.

Schoolchildren aged 6-7 years from two public schools in Piracicaba, SP, Brazil, were selected by a dentist, according to their caries activity. Those with cavitated carious lesions and/or active initial lesions (IL) were chosen. The dentists used mirror, CPI probe, air-drying for examination after the children brushed their teeth. The exclusion criteria were: use of fixed orthodontic device, presence of severe fluorosis and/or hypoplasias, and severe systemic diseases. For each training or calibration period of training, 10 to 13 different children were selected.

Caries Index, Codes and Criteria Adopted for the Study

Two diagnostic thresholds were used to record dental caries: 1) WHO threshold (DMFT index) following the WHO codes and criteria[17], in which a tooth is considered as decayed when a cavitation is present; 2) WHO+IL threshold, in which active initial lesions were also recorded following criteria adapted from Nyvad, et al.[13] and Fyfee, et al.[6]. An IL was defined as an active carious lesion which, upon visual assessment by a calibrated examiner, presented intact surface, no clinically detectable loss of dental tissue, with a rough, whitish/yellowish colored area of increased opacity presumed to be carious (when the CPI probe was used, its tip should be moved gently across the surface). IL locates close to gingival margin in smooth surfaces or extending along walls of fissure in occlusal surfaces. ILs detected in sealed, restored or cavitated surfaces were also recorded.

Training and Calibration of the Examiners

Each examiner was helped by a recorder during the study. A benchmaker examiner ("Gold Standard") conducted the training processes with both theoretical and practical activities, which lasted 20 hours; and the calibration exercises, which lasted 8 hours at each phase: baseline, 3 and 6 months after initial training. The training and calibration processes were conducted by the gold-standard examiner using two diagnostic thresholds on dental caries, as described in previous studies[2,3].

During theoretical discussions, the benchmaker examiner showed the examiners some photographic slides with clinical examples of each criterion that would be used in the study, in order to determine the examiners' knowledge about epidemiological diagnosis, to instruct them on the criteria and examination method to be used, and finally, to achieve an initial standardization among them. The mean interexaminer agreement obtained in this activity was Kappa=0.86.

The clinical training consisted of 4 periods of 4 hours each, and was conducted in an outdoor setting. Each dentist examined 10 to 13 children, with distinct caries activity and prevalence, *per* period. During this phase, examiners discussed clinical diagnosis, study codes and criteria, recording and other errors in order to reach an acceptable level of agreement (Kappa>0.85[17]).

The calibration exercises, in which the examiners did not discuss their findings, were carried out in 2 periods of 4 hours each, with a 1-week interval. These were also undertaken after 3 and 6 months, after the first calibration phase (baseline).

Conditions for the Examinations

The epidemiological examinations were carried out in an outdoor setting, under conditions such as natural light, with dental mirror and ball-ended CPITN probes with a diameter of 0.5 mm (to remove debris, assess the presence of fissure sealants and, in case of doubt, to check the surface texture of IL). Toothbrushing with fluoridated dentifrice was performed by the children, under supervision of a dental hygienist, following the Bass modified technique during approximately two minutes. Tooth air-drying (approximately 5 seconds *per* tooth) was performed by using compressed air delivered by a dental compressor (Wetzel: medical line 3.6/30 – 0.5 HP).

Data Analysis

A program using Microsoft Excel has been developed by the Department of Community Dentistry of FOP-UNICAMP to calculate the interexaminer reliability by means of the Kappa statistics[7] that has been recommended by the WHO[17] and the British Association of Community Dentistry[15] for evaluation of agreement among examiners in oral health surveys. The code recorded for each dental unit

or surface was entered for each examiner, in accordance with the different diagnosis thresholds (WHO; WHO+IL) used in the three calibration phases.

The objective of the statistical analysis was to assess interexaminer reliability under different caries diagnostic thresholds and different ways of analysis. Thus, interexaminer reliability was calculated by using Kappa statistics, according to two diagnosis thresholds (WHO; WHO+IL) and considering: a) the entire dentition; b) upper/lower jaws; c) sextants: upper/lower right/left, upper/lower anterior; d) each teeth individually. For each evaluation, codes from examination made by each examiner were compared to those of other examiners, (example: 1x2, 1x3… 1x11; 2x3… 2x11; 10x11). Average values of Kappa and its intervals of variation were calculated. Values above 0.85 were considered a high interexaminer reliability[17].

RESULTS

For both diagnostic thresholds, high mean values of Kappa were obtained (Tables). Moreover, interexaminer agreement was constant when considering the entire dentition, jaws and sextants (Tables 1 and 2).

Kappa values above 0.85 were obtained by analysis of sextants according to WHO threshold. However, when considering the WHO+IL threshold, the values for posterior sextants decreased (Tables 1 and 2).

The results of interexaminer reliability considering each tooth individually showed that the main difficulty was related to caries diagnosis in posterior teeth, especially the permanent first molars, for both thresholds with, in general, lower values for the WHO+IL threshold (Tables 3 and 4).

DISCUSSION

The process of examiners' calibration is an important aspect in planning and conducting oral health surveys. Brazilian surveys[4,10] have shown that training and calibration of examiners have been an aspect of great concern for measuring interexaminer agreement, according to the recommendations of the WHO, which has indicated the use of Kappa statistics[17]. Kappa test provides a better evaluation of disagreement among examiners during calibration processes since it is a measurement of adjusted agreement by taking into consideration the ratio of chance agreement[5]. Analysis of variance and post-hoc tests, such as Scheffé, have also been used to assess significant differences in caries indices among examiners[9].

The present study showed, in general, high means of interexaminer reliability for both diagnosis thresholds when considering the entire dentition, the upper/lower jaws and sextants (Tables 1 and 2).

On the other hand, lower Kappa values were observed for dental units (each tooth individually), especially when considering the most sensitive diagnosis threshold (WHO+IL) (Tables 3 and 4). As a consequence, good interexaminer agreement for the entire dentition may not be as real as if one considers separately the posterior teeth, in which cavitated and non-cavitated carious lesions are concentrated[8]. Moreover, considering the current epidemiological profile of dental caries, the higher number of sound teeth (fewer errors in diagnosis) in comparison to carious teeth (more errors in diagnosis) may dilute the errors attributed to carious teeth, leading to a positive vision of the results achieved in examiners' calibration[14]. Therefore, one may speculate that the analysis of Kappa values considering the entire dentition rather than each tooth individually may not be the best way to evaluate interexaminer reliability, especially in areas with low caries prevalence. It may be suggested the need for future reformulations in conducting examiners' calibration, paying

TABLE 1- Mean Kappa values for interexaminer reliability (interval of variation) under the WHO caries diagnosis threshold, considering the entire dentition, jaws and sextants according to calibration exercises. Piracicaba, 2004

	Calibration Exercises		
	Exercise 1 (Baseline)	Exercise 2 (3 months)	Exercise 3 (6 months)
Entire dentition	0.95 (0.93-0.99)	0.96 (0.93-0.99)	0.96 (0.94-0.98)
Upper jaw	0.96 (0.93-0.98)	0.95 (0.92-0.98)	0.95 (0.92-0.96)
Lower jaw	0.96 (0.93-0.97)	0.96 (0.95-0.97)	0.96 (0.94-0.97)
Upper right sextant	0.95 (0.92-0.97)	0.96 (0.92-0.98)	0.97 (0.96-0.98)
Upper anterior sextant	0.98 (0.97-1.00)	1.00 (0.97-1.00)	0.97 (0.97-1.00)
Upper left sextant	0.94 (0.90-0.95)	0.90 (0.87-0.96)	0.90 (0.94-0.92)
Lower left sextant	0.93 (0.89-0.95)	0.94 (0.89-0.96)	0.90 (0.91-0.96)
Lower anterior sextant	1.00 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.98-0.99)
Lower right sextant	0.96 (0.93-0.98)	0.96 (0.94-0.98)	0.91 (0.87-0.94)

more attention to diagnosis of posterior teeth and selecting children of distinct caries activity and prevalence[1].

Low Kappa values under the WHO+IL threshold can also be explained by the inherent difficulty in diagnosing IL, especially in surveys[1]. Although the examinations were carried out in sunny days under high luminosity conditions, the use of artificial light could generate an increase in interexaminer agreement by facilitating the view of posterior teeth. Further studies are needed to determine the relevance of using artificial light in dental caries surveys, mainly for detecting initial lesions.

In general, the use of more detailed measures to determine interexaminer agreement, such as the evaluation by dental unit (each tooth individually), improves the calibration process by showing which teeth are leading to great disagreements and indicating the possible need for greater efforts in training examiners[14].

However, it must be emphasized that the method of evaluating interexaminer agreement also depends on the study design and objectives, the desired degree of accuracy and the available resources. As an example, the calibration process using more rigorous statistical measures, such as the analysis by dental units, would be indicated in either clinical trials or case-control studies, in which the effect of preventive measures on the reduction of caries levels, including the detection of initial lesions, must be evaluated. Therefore the Kappa statistics considering reliability values according to each code/clinical condition can be employed[12].

TABLE 2- Mean Kappa values for interexaminer reliability (interval of variation) under the WHO+IL caries diagnosis threshold, considering the entire dentition, jaws and sextants according to calibration exercises. Piracicaba, 2004

	Calibration Exercises		
	Exercise 1 (Baseline)	**Exercise 2 (3 months)**	**Exercise 3 (6 months)**
Entire dentition	0.90 (0.85-0.96)	0.91 (0.85-0.98)	0.93 (0.88-0.96)
Upper jaw	0.91 (0.86-0.96)	0.91 (0.87-0.97)	0.92 (0.89-0.93)
Lower jaw	0.89 (0.85-0.96)	0.91 (0.86-0.97)	0.93 (0.90-0.93)
Upper right sextant	0.87 (0.83-0.88)	0.92 (0.82-0.95)	0.89 (0.88-0.96)
Upper anterior sextant	0.97 (0.95-0.98)	0.95 (0.94-0.99)	0.95 (0.92-0.97)
Upper left sextant	0.89 (0.82-0.90)	0.89 (0.79-0.96)	0.86 (0.82-0.89)
Lower left sextant	0.83 (0.80-0.91)	0.88 (0.77-0.94)	0.88 (0.84-0.90)
Lower anterior sextant	1.00 (0.99-1.00)	1.00 (1.00-1.00)	0.98 (0.98-0.99)
Lower right sextant	0.83 (0.80-0.85)	0.84 (0.78-0.98)	0.84 (0.79-0.87)

TABLE 3- Mean Kappa values for interexaminer reliability (interval of variation) under the WHO caries diagnosis threshold, according to calibration exercises and teeth. Piracicaba, 2004

Tooth	E1	E2	E3	Tooth	E1	E2	E3
16	0.72	0.88	0.83	36	0.60	0.96	0.86
55	0.90	0.86	0.74	75	0.89	0.88	0.71
54	0.90	1.00	0.84	74	0.89	0.84	0.85
53	0.95	1.00	0.91	73	1.00	0.80	1.00
52/12	1.00	1.00	0.97	72/32	1.00	1.00	1.00
51/11	0.97	0.89	1.00	71/31	1.00	1.00	1.00
61/21	0.96	0.71	1.00	81/41	1.00	1.00	1.00
62/22	1.00	1.00	0.97	82/42	1.00	1.00	1.00
63/23	1.00	1.00	1.00	83	1.00	1.00	0.87
64	0.97	0.86	0.76	84	0.98	0.89	0.97
65	0.83	0.59	0.86	85	0.85	0.86	0.87
26	0.80	0.69	0.78	46	0.77	0.89	0.86

E1=Exercise at baseline; E2= Exercise after 3 months; E3= Exercise after 6 months.

TABLE 4- Mean Kappa values for interexaminer reliability (interval of variation) under the WHO+IL caries diagnosis threshold, according to calibration exercises and teeth. Piracicaba, 2004

Tooth	E1	E2	E3	Tooth	E1	E2	E3
16	0.55	0.61	0.56	36	0.51	0.89	0.84
55	0.78	0.86	0.64	75	0.72	0.75	0.71
54	0.78	0.85	0.79	74	0.79	0.67	0.73
53	0.90	0.58	0.86	73	1.00	0.73	1.00
52/12	1.00	1.00	0.93	72/32	1.00	1.00	1.00
51/11	0.98	0.89	0.96	71/31	1.00	1.00	1.00
61/21	0.95	0.71	1.00	81/41	1.00	1.00	1.00
62/22	0.98	0.97	0.97	82/42	1.00	1.00	1.00
63/23	1.00	0.53	0.91	83	1.00	1.00	0.87
64	0.85	0.72	0.64	84	0.78	0.71	0.87
65	0.78	0.53	0.84	85	0.63	0.71	0.81
26	0.69	0.55	0.76	46	0.52	0.74	0.74

E1=Exercise at baseline; E2= Exercise after 3 months; E3= Exercise after 6 months.

On the other hand, in order to know and evaluate the epidemiological profile of dental caries in an underprivileged community, for instance, these more robust measures could be dispensed.

It is important to mention that the present study, which is part of a 12-month longitudinal study on examiners' calibration, presents some limitations, such as lack of validity results by comparing the examiners' results to those from the gold-standard examiner, and lack of intraexaminer errors. Such measures were not taken into consideration because the main goal of this study was to evaluate interexaminer reliability by using different diagnostic thresholds for caries detection as well as to verify the behavior of Kappa statistics in order to indicate the most adequate way to measure reliability during the examiners' calibration process in dental caries surveys.

CONCLUSION

The results of this study showed that the interexaminer reliability and its maintenance for six months were possible, under both caries diagnosis thresholds. Nevertheless, great disagreement was observed for the posterior teeth, especially when the WHO+IL criteria were used. The analysis considering dental elements individually was the best way of detecting disagreements among examiners during the calibration sections.

ACKNOWLEDGEMENTS

The authors are thankful to the School Principals, teachers, schoolchildren, and dentists for their invaluable participation in this study, as well as to CAPES (Coordination for the Improvement of Higher Education Personnel) for the scholarship granted to the first author during her Doctorate Course in Dentistry at FOP/UNICAMP.

REFERENCES

1- Assaf AV, Meneghim MC, Zanin L, Mialhe FL, Pereira AC, Ambrosano GMB. Assessment of different methods for diagnosing dental caries in epidemiological surveys. Community Dent Oral Epidemiol. 2004;32:418-25.

2- Assaf AV, Meneghim MC, Zanin L, Cortellazzi KL, Pereira, AC, Ambrosano, GMB. Effect of different diagnostic thresholds on dental caries calibration. J Public Health Dentistry. 2006;66:17-22.

3- Assaf AV, Meneghim MC, Zanin L, Tengan C, Pereira AC. Effect of different diagnostic thresholds for dental caries calibration - a 12 month evaluation. Community Dent Oral Epidemiol. 2006;34:213-9.

4- Brasil. Ministério da Saúde. SB Brasil 2003 - Projeto. Condições de saúde bucal da população brasileira: 2002-2003. Brasília, DF; 2004.

5- Cohen JA. Coefficient of agreement for nominal scales. Education and Psychological Measurement. 1960;20:37-46.

6- Fyffe HE, Deery C, Nugent ZJ, Nuttall NM, Pitts NB. Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee Selectable Threshold Method for caries diagnosis (DSTM). Community Dent Oral Epidemiol. 2000;28:42-51.

7- Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

8- Marthaler TM. Changes in dental caries 1953-2003. Caries Res. 2004 38:173-81.

9- Mitropoulos CM, Lennon MA, Worthington HV. A national calibration exercise for the British Association for the Study of Community Dentistry regional examiners. Community Dent Health. 1990;7:179-87.

10- Ministério da Saúde. Levantamento epidemiológico em saúde bucal: Brasil, Zona Urbana, 1986. Brasília: Centro de Documentação, Ministério da Saúde; 1988.

11- Nuttall NM, Paul JW. The analysis of inter-dentist agreement in caries prevalence studies. Community Dent Health. 1985;2:123-8.

12- Nyvad B, Machiulskiene V, Baelum V. Construct and predictive validity of clinical caries diagnostic criteria assessing lesion activity. J Dent Res. 2003;82:117-22.

13- Nyvad B, Machiulskiene V, Baelum V. Reliability of a new caries diagnostic system differentiating between active and inactive caries lesions. Caries Res. 1999;33:252-60.

14- Peres MA, Traebert J, Marcenes W. Calibração de examinadores para estudos epidemiológicos de cárie dentária. Cad Saude Publica. 2001;17:153-9.

15- Pine CM, Pitts NB, Nugent ZJ. British Association for the Study of Community Dentistry (BASCD) guidance on the statistical aspects of training and calibration of examiners for surveys of child dental health. A BASCD coordinated dental epidemiology programme quality standard. Community Dent Health. 1997;14(Suppl 1):18-29.

16- Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85:257-68.

17- World Health Organization. Oral health surveys. Basic methods. 4th ed. Geneva: WHO; 1997.